



Coding and Analyzing OPEN-ENDED QUESTIONS

»» Mine the Full Value of Open-Ended Questions

Survey questionnaires typically contain two broad types of questions: open-ended and closed-ended. Closed-ended questions present a discrete set of responses from which to choose. Such types of responses are easily quantified and analyzed while open-ended questions allow the respondent to answer a question in his own words. Such types of unstructured responses often provide richer and more valued information than closed-ended questions and are an important source of insight since they can generate information that was not anticipated. Despite their added value, researchers often prefer to avoid including open-ended questions in their surveys because of the tedious task of reading and coding responses, a time-consuming and expensive task, especially when one has more than a few hundred written responses.

QDA Miner and **WordStat** represent efficient alternatives to such time-consuming procedures by offering unique computer assistance for coding and analyzing open-ended responses in a fraction of the time normally required when such a task is performed manually. We will review some of those features and provide some resources to choose the best strategy and to get started. While we focus here on the coding and analysis of open-ended questions from survey data, the same techniques could also be applied for other types of data consisting of a large number of short text responses, such as customer feedback or, Twitter and other social media feeds.

» QDA Miner or WordStat?

Both QDA Miner and WordStat offer useful tools for coding and analyzing open-ended questions. The answer to the question about which software to use depends on a variety of factors, the most obvious one being the quantity of responses one has to code. Such a decision also depends on other, external factors such as the available time and financial resources to code and analyze these responses as well as whether the data comes from a one-time survey, from a recurring survey or an ongoing data-collection process.

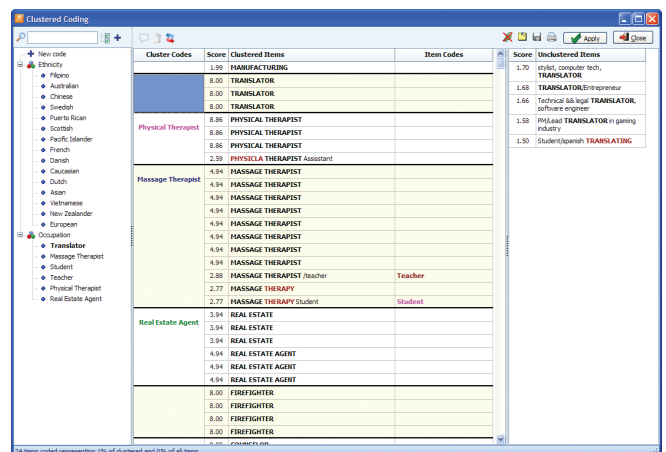
When to Use QDA Miner

QDA Miner is essentially a manual coding tool where the final decision whether to apply a code to a specific response remains under the control of a human coder. While QDA Miner offers numerous computer-assistance features that can help achieve faster and more reliable coding, the coder always has the opportunity to review the suggestions before any coding is applied. The most basic computer-assistance feature, and the one that people are the most familiar with, is the text-search tool that allows one to retrieve and code all responses containing specific key words or key phrases. QDA Miner provides such a tool, allowing one to use complex search expressions with Boolean operators (AND, OR, NOT) as well as thesaurus-based searches. However, the software goes much beyond this by offering advanced search functions relying on information-retrieval techniques such as machine learning, relevance feedback, fuzzy string matching, etc. Here are some of the most useful advanced tools for analyzing transcripts of open-ended questions:

- The **Cluster Retrieval** tool is a truly innovative coding device relying on unsupervised machine learning. It automatically groups similar responses (even with spelling mistakes), and allows the coder to review and quickly code those clusters using an intuitive drag-and-drop coding device. It typically speeds up the manual coding of open-ended responses by a factor that varies between three and 100 times faster than what it would take to manually code similar unclustered written data.
- The **Code Similarity** feature allows one to quickly identify text segments similar to items that have been previously coded either in the current project or in another project. This feature could be used to speed up the coding of partially coded projects. It may also be used on fully coded projects to identify items that may have been missed. This feature can be used to increase both the speed and the reliability of the coding.

- The **Query-by-Example** retrieval tool allows the coder to focus on a specific text segment, or all responses associated with a specific code, and to retrieve other responses that share some similarities. It includes a “relevance feedback” feature that allows the coder to indicate, from the list of retrieved text segments, those that are relevant or irrelevant, causing the software to learn from that feedback, refine the search, and retrieve more relevant examples.

Once the responses have been coded, QDA Miner will offer numerous tools to retrieve all responses associated with a specific code or a combination of codes, compute descriptive statistics and create presentation graphics on those codes (bar charts, pie charts, tag clouds), examine their co-occurrences using cluster analysis, multidimensional scaling or proximity plots, as well as relate codes associated with specific open-ended responses with those from closed-ended questions using tools like crosstabulation, correspondence analysis, heatmaps, bubble charts, and more. QDA Miner also integrates an inter-raters reliability feature to test the level of agreement among coders.



The powerful cluster extraction and coding tool allows one to group similar sentences or paragraphs into clusters.

When to Use WordStat

WordStat provides a very different solution for the coding and analysis of answers to open-ended questions. Rather than relying on human coders, the software offers numerous tools to partly or fully automate the analysis of large number of responses or to quickly identify themes and patterns in text data without the necessity of reading individual responses. WordStat is especially appropriate when analyzing very large numbers of open-ended responses (from several hundreds to tens or hundreds of thousands of responses), or when one wants to develop a coding solution that may later be reapplied to similar text responses. While a manual revision is still possible, such a revision typically occurs after the text data has been coded or classified. WordStat offers three broad types of strategies to analyze responses to open-ended questions:

1. TEXT MINING - One can adopt an exploratory approach to text data by applying a combination of NLP (natural language processing) techniques and statistical methods. Themes in open-ended may be extracted automatically using topic modeling (Latent Semantic Analysis), or by applying Hierarchical Clustering Techniques to analyze the co-occurrence of the most common words and phrases. The Proximity Plot may then be used to show words or phrases that are most often associated with a specific topic, brand, person or company name and to quickly compare those with the ones associated with another target name or topic. Correspondence Analysis may also be used to identify patterns of words, phrases, or topics that are specific to different groups of respondents. All these exploratory techniques can produce insightful results from a large collection of textual data in a matter of seconds.

2. CONTENT ANALYSIS - WordStat offers state-of-the-art content analysis tools that may be used to build and apply categorization dictionaries (what others may call "taxonomies"). The main idea behind this approach is to measure references to specific concepts or themes by identifying the various ways one could express such ideas. A content-analysis dictionary can consist of a large number of categories, where each category may itself contain hundreds of words, word patterns, phrases and rules. One can apply existing content-analysis dictionaries developed by others or create one's own dictionary. While developing and validating a dictionary does require some time, when analyzing large data sets consisting of a thousand or more open-ended responses, such an approach will typically be faster than manually coding individual responses. Furthermore, once developed, a content-analysis dictionary can be applied on new data sets consisting of

responses to similar questions and can produce results in a matter of seconds. Such an approach is thus ideal if one needs to partly automate the coding of text responses from recurring surveys or for analyzing ongoing text-data collections such as text from social-media or customer-feedback systems.

3. AUTOMATIC DOCUMENT CLASSIFICATION - A third approach consists of coding open-ended responses by applying supervised machine-learning techniques to automatically classify responses into one or several categories. Such an approach requires the availability of a relatively large number of responses already coded. The coding of those responses are used as learning examples by which the software will attempt to automatically identify similar examples of those in uncoded responses using key words or phrases. WordStat offers a choice between two popular machine-learning algorithms: Naive Bayes and k-Nearest Neighbors. The automatic classification technique may be used to generate dichotomous decisions (present or absent), a nominal classification (one out of "n" mutually exclusive values) or to generate a score on an ordinal scale (such as a Likert scale).

While each of the above techniques may be used separately, one will often profit from combining these. For example, a hierarchical clustering of the most frequent words may be used to identify the most common topics as well as an organizing structure of those topics. Such a structure may then be replicated in a content-analysis dictionary by the creation of specific content categories to measure those themes and by grouping those specific categories into broader content categories. A correspondence analysis on words may also be used to identify hypotheses that may later be tested by the development of more comprehensive dictionaries. One may also use a content-analysis dictionary to confine the automatic document classification to the use of specific keywords or to make sure close synonyms or inflected or misspelled forms of words will not be ignored or be treated independently from each other.



» Why Use Both QDA Miner and WordStat

There are many ways one can combine the computer-assisted qualitative coding features of QDA Miner with the content-analysis and text-mining tools available in WordStat. Even those who rely entirely on manual coding will find WordStat exploratory text analysis features useful for familiarizing themselves with the range of topics being mentioned and to identify not only potential codes but also a codebook structure that will be appropriate for organizing all those codes. WordStat content analysis or automatic document classification features may also be useful to select a limited set of potentially more relevant responses from a huge number of text responses (too large to be coded manually). These more relevant responses may then be examined more carefully by human coders.

Those researchers relying on WordStat for automatically quantifying text responses will often feel the need to review the results and to make manual adjustments in order to achieve greater precision. They may also find some topics difficult to identify automatically and may thus need to go back to a more manual approach of coding. Finally, the development of an automatic document classification model requires the availability of a training set of responses carefully categorized by human coders. If such a training set is not available, the researcher will have no choice but to create one. QDA Miner computer-assistance features are very useful for quickly coding and validating such a training set.

Some Additional Resources

The following web page offers additional resources to help you learn a little bit more about using QDA Miner and WordStat for analyzing open-ended questions from surveys or from customer-feedback questionnaires:

www.provalisresearch.com/Open-Ended

Surveys using QDA Miner or WordStat

The following list is studies from researchers who have used QDA Miner, WordStat or both for analyzing responses to open-ended questions. The size of these surveys varies a lot – from a small sample of 104 environmental scientists (Wright & Wyatt, 2008) to a much larger survey of more than 41,500 federal government employees.

- Ames, D. R. & Bianchi, E.C. (2008). The agreeableness asymmetry in first impressions: Perceivers' impulse to (mis)judge agreeableness and how it is moderated by power. *Personality and Social Psychology Bulletin*, 34, 1719-1736.
- Behruzi, R., Hatem, M., Goulet, L., Fraser, W. (2011). The facilitating factors and barriers encountered in the adoption of a humanized birth care approach in a highly specialized university affiliated hospital. *BMC Women's Health*, 11(1):53 (2011) PMID 22114870.
- Frank, B.A. & Walsh, R.J. (2011). Does reflective learning take place in online MBA introductory quantitative courses? *Journal of Instructional Pedagogies*.
- Harper, C.A., Shaw, C.E., Fly, J.M., & Beaver, J.T., (2012). Attitudes and motivations of Tennessee deer hunters toward quality deer management. *Wildlife Society Bulletin*, 36(2), 277-285.
- Kim, S., Cha, J., Cichy, R.F., Kim, R., & Tkach, J.L. (2016). Roles of private club volunteer leaders: An exploratory study of content analysis. *International Journal of Hospitality & Tourism Administration*, 17(1), 43-71.
- Latham, S. (2009). Contrasting strategic response to economic recession in start-up versus established software firms. *Journal of Small Business Management*, 47(2), 180-201.
- Leurer, M. D., & Misskey, E. (2015). "Be positive as well as realistic": A qualitative description analysis of information gaps experienced by breastfeeding mothers. *International Breastfeeding Journal*, 10(1), 1-11.
- McCandless, D.W, Bruce, R.D., Gebken, R.J. (2010). Retention of Construction teachers in secondary education. *International Journal of Construction Education and Research*, 6(2), 104-121.
- McComas, K.A., Besley, J.C., & Trumbo, C.W. (2007). Why citizens do and do not attend public meetings about local cancer cluster investigations. *Policy Studies Journal*, 34(4), 671-698.
- Pullman, M. McGuire, K, Cleveland, C. (2005). Let me count the words: Quantifying open ended interactions with guests. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3), 323-343.
- Saunders, M.N.K. (October, 2011). Web versus mail: The influence of survey distribution mode on employees' response. *Field Methods*. Online. October 9, 2011.
- Spinks, N., Silburn, N., & Birchall, D. (2006). Educating engineers for the 21st century: The industry view. Henly Management College: Henly on Thames: UK.
- U.S. Merit Systems Protection Board (2011). Making the Right Connections: Targeting the Best Competencies for Training. Washington, D.C.
- Wright, T.S.A, & Wyatt, S.L (2008). Examining influences on environmental concern and career choice among a cohort of environmental scientists. *Applied Environmental Education and Communication*, 7, 30-39.